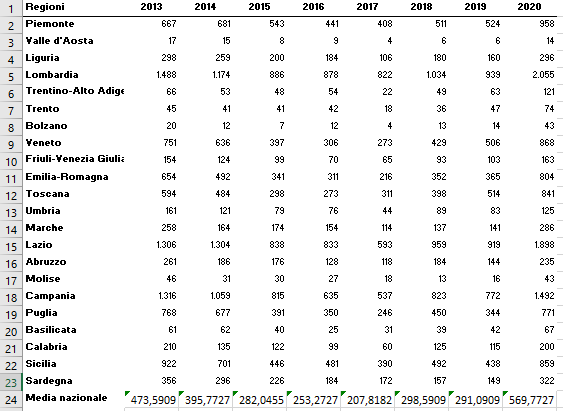
# Introduzione

Si sente spesso parlare di violenza di genere, ma che cosa si intende? Le Nazioni Unite la definiscono come “*ogni atto legato alla differenza di sesso che provochi o possa provocare un danno fisico, sessuale, psicologico o una sofferenza della donna, compresa la minaccia di tali atti, la coercizione o l’arbitraria privazione della libertà sia nella vita pubblica che nella vita privata”* [1]*.* Contro la violenza di genere è stato realizzato ed è diventato di fondamentale importanza il numero verde 1522 attivo 24 ore su 24 tutti i giorni dell’anno che offre accoglienza in italiano, inglese, francese, spagnolo e arabo [2].

## Caso di studio

Nel 2020 a causa della pandemia è stato vissuto il lockdown. Si è sentito molto parlare degli effetti economici che questo ha causato ma poco degli effetti sociali e, in particolare, sulla violenza di genere. Si è deciso pertanto di analizzare i dati relativi alle chiamate al numero antiviolenza e antistalking 1522 effettuate nel periodo marzo-giugno, confrontandole con quelle dello stesso periodo negli anni precedenti. Secondo un rapporto pubblicato dall’Istat, tale numero è più che raddoppiato nel 2020 rispetto allo stesso periodo del 2019. Nella seguente relazioni si prenderanno in considerazione due tabelle di dati che rappresentano, rispettivamente, gli utenti e le vittime al numero 1522 effettuate nel periodo marzo-giugno suddivise per regione ed anno (2013-2020). In particolare, verranno esaminate nei dettagli le curve relativi ai dati della regione Campania e la media delle chiamate effettuate sull’intero territorio nazionale.

Di seguito vengono riportate le due tabelle di dati.



## Rappresentazione grafica

Sono stati creati due barplot, uno relativo ai dati della media nazionale e l’altro relativo ai dati in Campania.

# Statistica descrittiva univariata

In questo capitolo verranno sviluppate le analisi della statistica univariata per i campioni di dati della Campania e della media nazionale. Prima di tutto verrà mostrata la funzione di distribuzione di distribuzione empirica continua, successivamente verranno mostrati i valori degli indici di sintesi di posizione e dispersione. Infine, verranno mostrati gli indici di sintesi che consentono di studiare la forma della distribuzione di frequenze.

## Funzione di distribuzione empirica continua

La funzione di distribuzione empirica continua viene utilizzata per fenomeni quantitativi continui e si tratta di una funzione di distribuzione strutturata in classi. Supponiamo di voler dividere i dati in k classi C1 = [z0, z1), C2 =[z1, z2), … Ck =[zk-1, zk], con z0 < z1 < z2 < … zk-1 < zk, dove z0 corrisponderà al minimo delle osservazioni, mentre zk corrisponderà al massimo delle osservazioni. La funzione di distribuzione empirica continua verrà calcolata a partire dalle frequenze relative comulative delle varie classi.

### Funzione di distribuzione empirica continua Utenti

Per calcolare la funzione di distribuzione continua sono stati suddivisi le osservazioni sono state suddivise in tre classi. Per quanto riguarda la media nazionale le classi individuate sono le seguenti:

C1 = [208, 329), C2 = [329, 450), C3 = [450, 570]. Per quanto riguarda la Campania le classi individuate sono le seguenti: C1 = [537, 855), C2 = [855, 1173), C3 = [1173, 1492]. Sono stati quindi creati i grafici che mostrano le frequenze di distribuzione continue della Campania e dell’intera nazione.





### Funzione di distribuzione empirica continua Vittime

## Indici di sintesi

Alcuni indici di sintesi utili a descrivere i dati sono media, mediana, moda, varianza, deviazione standard e coefficiente di variazione. Le prime tre sono misure di centralità dei dati mentre le altre misurano la loro dispersione.

Supponiamo di avere un insieme x1, x2, .., xn di n valori numerici. Si definisce **media campionaria** e si denota con la quantità:

Dato un campione di dati di ampiezza n ordinato in maniera crescente, la **mediana campionaria** è l’indice che consente di dividere il campione a metà. Se n è dispari, si definisce mediana campionaria il valore in posizione (n+1)/2, se invece n è pari la mediana campionaria è definita come la media aritmetica dei valori nelle posizioni n/2 e n/2+1. Oltre alla mediana campionaria che consente di dividere l’insieme dei dati a metà vengono considerati anche altri indici di posizione, detti quantili. I **quantili** consentono di dividere l’insieme dei dati ordinato in un numero fissato di parti uguali, successivamente verranno considerati i **quartili** che consentono di dividere l’insieme dei dati in quattro parti uguali. La mediana corrisponde al secondo quartile.

La **moda campionaria** di un insieme di dati, se esiste, è la modalità a cui è associata la frequenza più elevata. Se esistono più modalità con frequenza massima, ciascuna di esse è detto valore modale. Per le variabili raggruppate in classi, è possibile individuare la classe modale che corrisponde alla classe con altezza massima nell’istogramma delle frequenze.

Assegnato un campione x1, x2, .., xn di dati numerici, si definisce **varianza campionaria** e si denota con s2 la quantità: , dove denota la media campionaria. Inoltre, si definisce deviazione standard campionaria la radice quadrata della varianza campionaria, ossia:

Assegnato un campione x1, x2, .., xn di dati numerici, si definisce coefficiente di variazione il rapporto tra la **deviazione standard campionaria** e il modulo della media campionaria, ossia:

Momenti

Skewness

curtosi

### Indici di sintesi Utenti

Nel grafico seguente vengono mostrate le due curve relative ai dati che si stanno analizzando.



Entrambe le curve mostrano una distribuzione di frequenze non simmetrica, in particolare inizialmente sono decrescenti ed hanno un picco massimo nell’ultimo anno 2020. Le due curve sono tra loro abbastanza simili.

Il grafico seguente mostra, invece, i boxplot di entrambi i dati per illustrare alcune caratteristiche della distribuzione di frequenza come centralità, dispersione, forma e la presenza di eventuali valori anomali. ….??



Entrambi i boxplot rivelano la presenza di asimmetria nei dati in quanto le distanze tra primo e terzo quartile dalla linea della mediana sono molto diverse tra loro. Si può intuire che le curve hanno una coda più allungata a destra e ciò verrà confermato attraverso il calcolo della skewness campionaria.

Utilizzando la funzione summary in R è possibile calcolare minimo, massimo, media, mediana, primo e terzo quartile. Dai grafici a barre in alto è possibile inoltre determinare la moda, che in entrambi i casi è associata all’anno 2020. In generale, la curva dei due dati è sostanzialmente simile anche se i dati relativi all’intera nazione sono più bassi in quanto sono calcolati dalla media di tutte le nazioni, che viene fortemente influenzata dai valori bassi presenti in molte regioni con meno abitanti rispetto alla Campania.





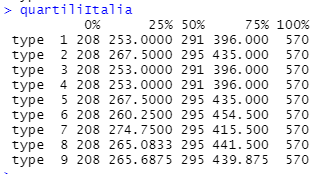
Per individuare la moda si considerano gli istogrammi delle frequenze dei dati considerando la loro suddivisione nelle seguenti classi: C1 = [0, 500), C2 = [500, 1000), C3 = [1000, 1500) C1 = [1500, 2000), C2=[2000, 2500].

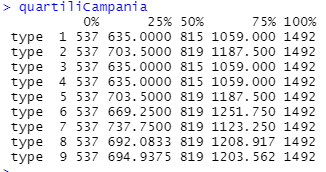
La classe modale per l’Italia è la prima, in particolare tutti i valori sono concentrati in quella classe. Per la Campania invece la classe modale è la seconda.





**Quartili con i differenti algoritmi di R.** In R esistono 9 differenti algoritmi per il calcolo dei quartili. Di seguito vengono mostrati i risultati ottenuti con ciascuno dei diversi algoritmi sul campione di dati della media nazionale e della Campania.





**Indici di dispersione**



### Indici di sintesi Vittime

## Forma della distribuzione di frequenze

**Skewness campionaria**. Entrambe le distribuzioni di frequenze hanno un’asimmetria positiva, la distribuzione di frequenza ha quindi una coda più allungata a destra.



**Curtosi campionaria.** Il valore di entrambe le curtosi campionarie è negativo quindi entrambe le distribuzioni di frequenza sono meno piccate di una distribuzione di frequenze normale standard.



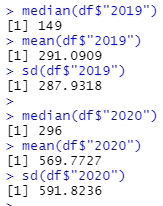
# Statistica descrittiva bivariata

## Regressione lineare semplice

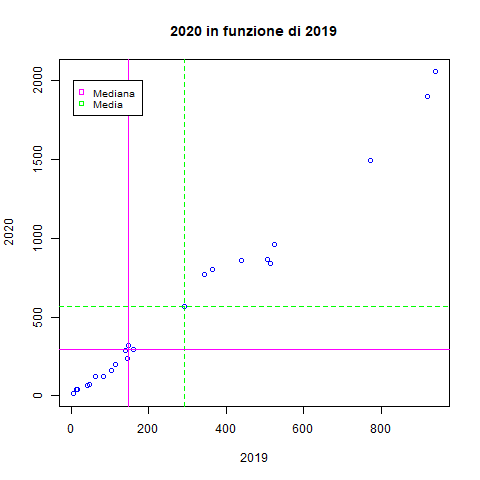
### Regressione lineare semplice Utenti

Le variabili che vengono considerate in quest’analisi sono le colonne della tabella relative ai dati del 2019 e del 2020. In particolare, la variabile indipendente è 2019, quella dipendente è 2020. Si calcolano i valori degli indici statistici mediana, media e deviazione standard dei dati relativi alle variabili considerate.

Si nota che sia mediana, sia media che deviazione standard sono maggiori per la seconda variabile.



Successivamente, si realizza lo scatterplot ponendo sulle ascisse la variabile indipendente 2019 e sulle ordinate la variabile dipendente 2020. Vengono poi tracciate delle linee orizzontali e verticali in corrispondenza delle mediane e delle medie delle due variabili.



Dallo scatterplot si nota che i dati sono posizionati lungo una retta ascendente quindi si può dedurre che esiste una correlazione positiva tra le variabili considerate.

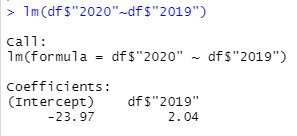
Per vedere se esiste tale correlazione si calcolano la covarianza e la correlazione campionaria. Da questo calcolo si evince che i dati dei due vettori 2019 e 2020 sono positivamente correlati essendo la covarianza positiva. Inoltre, il coefficiente di correlazione è uguale a 0.9923597 che è prossimo ad 1, quindi come indicato dallo scatterplot esiste una forte correlazione lineare tra i dati del 2019 e i dati del 2020.



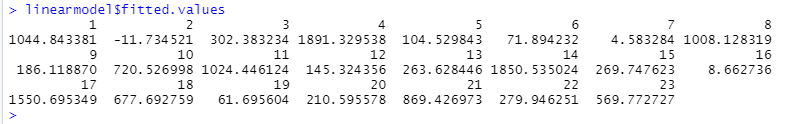
Il seguente grafico mostra lo scatterplot relativo ai dati del 2019 e del 2020 con la retta interpolante stimata.



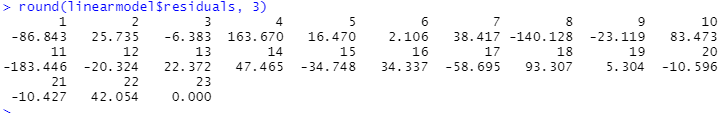
Il seguente codice permette di ottenere il modello di regressione lineare per le due variabili. In particolare, l’intercetta vale -23.97, mentre il coefficiente angolare vale 2.04. Siccome il coefficiente angolare è positivo, la retta è ascendente.



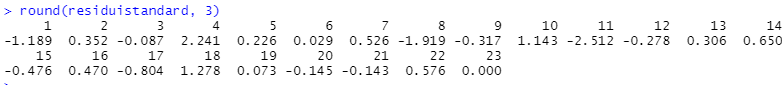
Il codice seguente permette di visualizzare i valori stimati.



Il seguente codice permette di visualizzare i residui, ossia di quanto i valori osservati si discostano dai valori stimati.



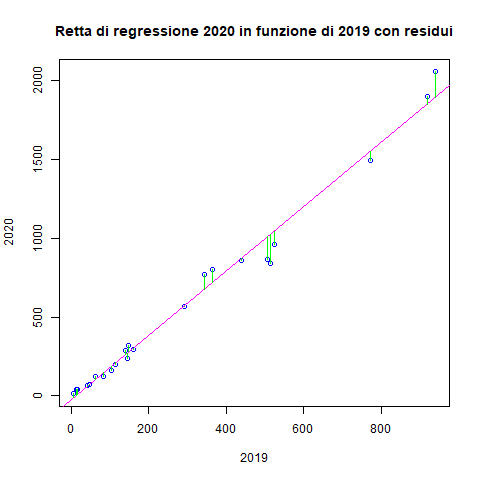
Valore dei residui standardizzati rispetto alla deviazione standard. Si può osservare che i valori sono molto piccoli.



Le seguenti linee di codice mostrano i valori della mediana, della varianza e della deviazione standard dei residui.







Per valutare quanto la retta di regressione si adatta ai dati si calcola il coefficiente di determinazione che si calcola effettuando il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati. In questo caso il coefficiente di correlazione vale 0.9848. Siccome è prossimo ad 1, significa che la retta descrive bene i dati considerati, infatti anche dai grafici visti precedentemente si nota che gli scostamenti dalla retta sono molto piccoli.



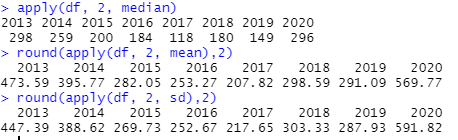
### Regressione lineare semplice Vittime

## Regressione lineare multipla

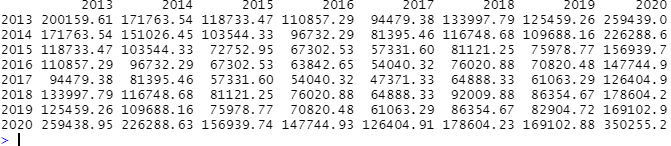
### Regressione lineare multipla Utenti

Si utilizza il modello di regressione lineare multipla per spiegare la relazione le variabili indipendenti: 2013, 2014, 2015, 2016, 2017, 2018, 2019 e la variabile dipendente: 2020

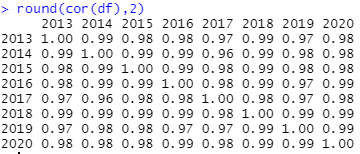
Valore degli indici di posizione e di dispersione (mediana, media e deviazione standard) relativi alle variabili:



Matrice delle covarianze:



La matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili.

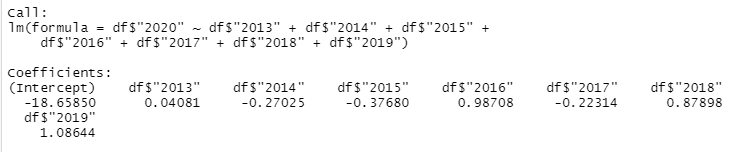


Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

Il seguente grafico visualizza in un’unica finestra tutti gli scatterplot ottenuti mettendo in relazione le varie coppie di variabili. Da tale grafico si può dedurre che le variabili sono altamente correlate e si intuisce che avranno un coefficiente di correlazione quasi pari ad 1.



Utilizzando il modello di regressione lineare multipla si ottiene:

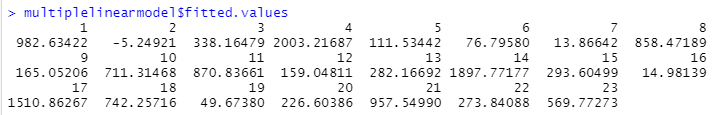


Da cui si ricava che l’intercetta è -18.65850 e i regressori sono: 0.04081, -0.27025, -0.37680, 0.98708, -0.22314, 0.87898, 1.08644.

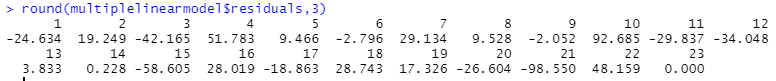
I segni dei regressori b1, b4, b6, b7 sono positivi: questo indica che all’aumentare del numero di utenti nel 2013, 2016, 2018 e 2019 aumenta il numero di utenti nel 2020. Mentre i regressori b2, b3, b5 sono negativi quindi all’aumentare del numero di utenti nel 2014, 2015, 2017 diminuisce il numero di utenti nel 2020.

Il regressore di 2013 è prossimo allo zero, questo indica che il numero di utenti nel 2013 non incide in maniera significativa il numero di utenti nel 2020.

Valori stimati rispetto al modello di regressione multipla.



Residui dei valori osservati rispetto ai valori stimati.



Valori dei residui standardizzati rispetto alla deviazione standard.

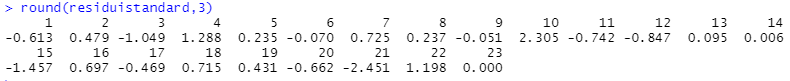
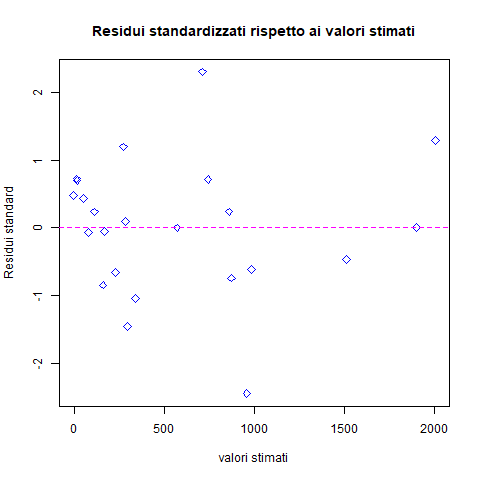


Grafico che mostra i residui standardizzati in funzione dei valori stimati.



I punti indicano dove si collocano i residui rispetto ai valori stimati. Non si evidenzia nessuna tendenza particolare rispetto alla retta orizzontale che rappresenta la media dei residui (0).

Anche in questo caso il coefficiente di determinazione è prossimo ad 1, infatti vale 0.9954. Il modello di regressione lineare multipla descrive bene i dati considerati.



### Regressione lineare multipla Vittime

# Analisi dei cluster

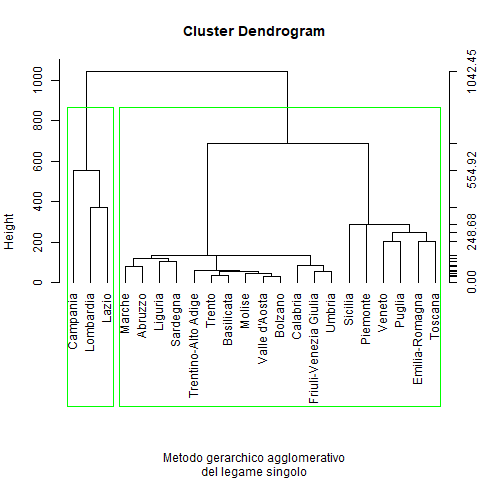
## Analisi dei cluster Utenti

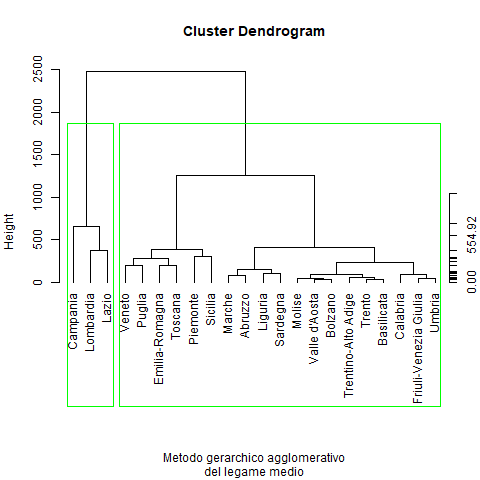
Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

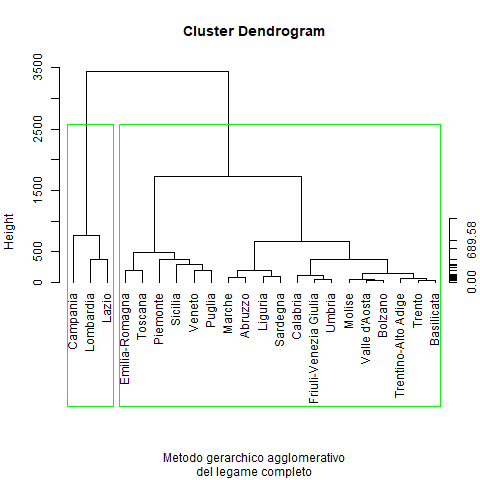
Primo cluster: 19 individui

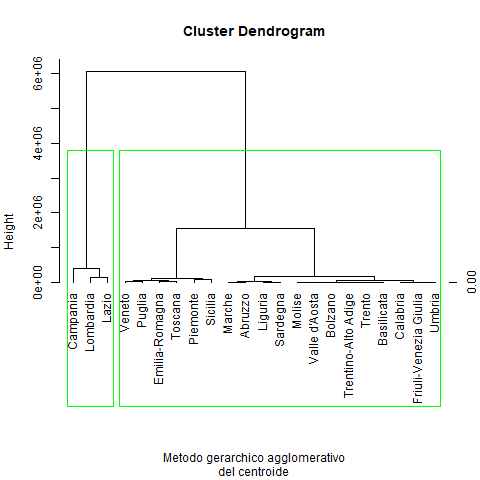
Secondo cluster: 3 individui

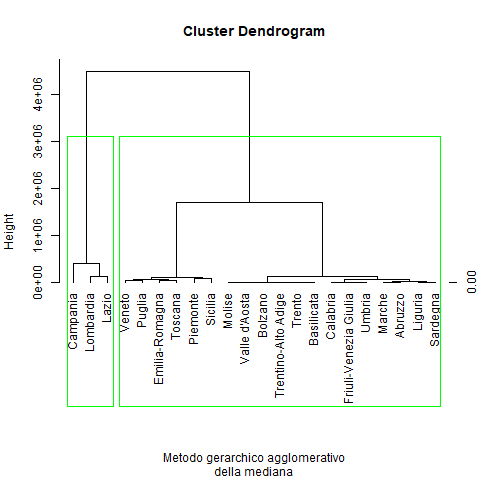
|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Molise, Puglia, Basilicata, Calabria, Sicilia, Sardegna |
| Cluster 2 | Lombardia, Lazio, Campania |











La misura di non omogeneità totale trT nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 7304986.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 340968.7.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2= 15681146.

Il rapporto =  **0.6722287**.

Il metodo non gerarchico K-means ha fornito il seguente partizionamento in due cluster.

Primo cluster: 9 individui

Secondo cluster: 13 individui

|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Lombardia, Veneto, Emilia-Romagna, Toscana, Lazio, Campania, Puglia, Sicilia |
| Cluster 2 | Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Friuli-Venezia Giulia, Umbria, Marche, Abruzzo, Molise, Basilicata, Calabria, Sardegna |

La misura di non omogeneità totale trT nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 5812676.7.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 637988.3.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2=16876436.

Il rapporto = **0.7234691**.

La suddivisione in cluster ottenuta con il metodo non gerarchico K-means risulta essere migliore.

## Analisi dei cluster Vittime

# Variabile aleatoria esponenziale

La **distribuzione esponenziale** è una distribuzione di probabilità continua che descrive la "durata di vita" di un fenomeno che *non invecchia*. Un esempio è la *durata di vita* di una particella radioattiva prima decadere oppure la durata della richiesta di un servizio. Si dice che X ha distribuzione esponenziale di parametro λ>0 e si indica con X∼EXP(λ), se la sua funzione di densità di probabilità è:

E corrispondente funzione di distribuzione:

Per una variabile esponenziale si ha che:

Osservando che , se X rappresenta un tempo allora λ rappresenta una frequenza. Quindi se la variabile aleatoria descrive, ad esempio, la durata di vita di un componente elettronico si intuisce che i tempi di vita maggiori corrispondono ai parametri λ più piccoli. Infatti, la funzione densità, al diminuire di λ, si schiaccia sull’asse delle ascisse. Di conseguenza la media si sposta verso valori più elevati e il componente si guasta mediamente più tardi. Pertanto, λ risulta essere inversamente proporzionale al tempo di vita medio del componente.

Come specificato precedentemente, tale variabile aleatoria descrive un fenomeno che non invecchia, ciò significa che è privo di memoria. Gode infatti della seguente proprietà di “assenza di memoria”, per ogni s, t reali positivi risulta:

Se si interpreta X come un tempo di attesa, la precedente equazione mostra che la probabilità condizionata che il tempo di attesa X sia maggiore di t+s dato che essa è maggiore di s non dipende da quanto si è già atteso, ossia da s.

Nel seguente grafico è rappresentata la densità di probabilità e la funzione di distribuzione di una variabile aleatoria con distribuzione esponenziale e parametro λ=3.



La probabilità che la variabile aleatoria esponenziale con λ=3 assuma valori nell’intervallo (0.5, 1.5) è 0.2120212.



I quartili sono:



**Confronto della densità teorica esponenziale con la densità simulata**. Si può notare che all’aumentare dell’ampiezza del campione, l’istogramma delle frequenze relative si avvicina alla densità esponenziale teorica.



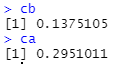
## Stima puntuale e stima intervallare

Si desidera studiare una popolazione descritta da una variabile aleatoria X con funzione di distribuzione esponenziale. In particolare, il campione ha ampiezza 50 e denota i tempi di gestione in minuti di una richiesta da parte di un servizio A. Si vuole stimare il parametro non noto λ.

La **stima puntuale** del parametro non noto con il metodo dei momenti e della massima verosimiglianza forniscono come stimatore la media campionaria. Risulta quindi λ = 0.1876024. La media campionaria è uno stimatore corretto con varianza minima e consistente per 1/ λ.

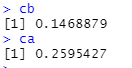
La **stima intervallare** si propone, a differenza della stima puntuale, di determinare in base ai dati del campione un limite superiore e un limite inferiore entro il quale sia compreso il parametro non noto λ con un certo coefficiente di confidenza (o grado di fiducia). Per effettuare la stima intervallare su un campione con distribuzione esponenziale viene utilizzato il teorema centrale di convergenza che necessita un campione di ampiezza ≥30. Siccome il campione a disposizione ha ampiezza pari a 50, risulta possibile applicare tale metodo.

**Stima dell’intervallo di confidenza di grado 1 − α = 0.99**. Il limite inferiore risulta cb=0.1375105, mentre il limite superiore risulta ca=0.2951011.



Risulta quindi:

**Stima dell’intervallo di confidenza di grado 1 − α = 0.95.** Il limite inferiore risulta cb=0.1468879, mentre il limite superiore risulta ca=0.2595427.

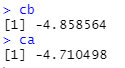


Risulta quindi:

Si nota che all’aumentare del grado di confidenza 1 – α l’intervallo diventa più grande.

## Confronto tra due popolazioni esponenziali

Si desidera confrontare i tempi per gestire una richiesta da parte di due servizi A e B. I tempi sono distribuiti come una variabile esponenziale. Il servizio A gestisce 50 richieste, mentre il servizio B ne gestisce 80 con i seguenti risultati sulle medie e sulle deviazioni standard dei tempi: mediaA=5.330421, sdA=4.737098, mediaB=10.11495, sdB=10.82139. Si vuole determinare una stima dell’intervallo di confidenza di grado 1 − α = 0.99 per la differenza 1/ λ A− 1/ λ B tra i tempi medi impiegati per gestire una richiesta da parte dei due servizi.



Risulta quindi:

Siccome ca e cb sono entrambi negativi, la differenza 1/ λ A − 1/ λ B risulta essere negativa, pertanto λ A > λ B . Siccome in una variabile aleatoria esponenziale λpuò essere visto come una frequenza, il servizio A è in grado di servire più richieste per minuto rispetto al servizio B.